Learning Bayesian Network Parameters from a Small Data Set: A Further Constrained Qualitatively Maximum a Posteriori Method

Abstract

To improve the learning accuracy of the parameters in a Bayesian network from a small data set, domain knowledge is normally incorporated into the learning process as parameter constraints. MAP-based (Maximum a Posteriori) methods that utilize both sample data and domain knowledge have been well studied in the literature. Among all the MAP-based methods, the QMAP (Qualitatively Maximum a Posteriori) method is one of the algorithms with the highest learning performance. When the data is insufficient, however, the estimation given by the QMAP often fails to satisfy all the parameter constraints, and this has made the overall QMAP estimation unreliable. To ensure that a QMAP estimation does not violate any given parameter constraints and further to improve the learning accuracy, a FC-QMAP (Further Constrained Qualitatively Maximum a Posteriori) algorithm is proposed in this paper. The algorithm regulates QMAP estimation by replacing data estimation with a further constrained estimation via convex optimization. Experiments and theoretical analysis show that the proposed algorithm outperforms most of the existing parameter learning methods including Maximum Likelihood, Constrained Maximum Likelihood, Maximum Entropy, Constrained Maximum Entropy, Maximum a Posteriori, and Qualitatively Maximum a Posteriori.

Keywords: Bayesian network, Parameter learning, Small data set, Domain knowledge

1. Introduction

A Bayesian network (BN) is typically a directed acyclic graph representing a model that combines probability theory and graphical model theory. The BN was systematically introduced in 1988[1] by Judea Pearl. Following an intensive research for about 30 years worldwide, BNs have become a powerful tool for uncertainty analysis and have been applied to deal with a wide range of issues, including gene analysis[2], robot control[3], fault diagnoses[4], target tracking[5], signal processing[6], ecosystem modeling[7] and educational measurement[8].

Usually, to construct a BN from data, reasonably-sized samples are required, depending on the topology and complexity of the network. If sufficient samples are available, the construction of an accurate BN is easy and can be accomplished by traditional methods such as Maximum Likelihood (ML)[9]. Unfortunately, however, collecting a large amount of data is difficult for decision-making problems under some certain circumstances, such as rare disease diagnosis[10], earthquake prediction[11] and parole assessment[12]. In such cases, knowledge from domain expert is often considered supplementary information in constructing the network.

In practice, domain experts usually feel more comfortable with providing qualitative parameter constraints in the form of $p_1 > 0.8$, $p_1 \approx p_2$, $p_1 > p_2$, $(p_1 + p_2) > (p_3 + p_4)$[13], etc., where p_1 and p_2 denote parameters in a Bayesian network. Those constraints look simple but can be very effective for improving the BN modeling accuracy, especially when the given data set is small. In this paper, we focus on learning BNs in cases where the sample size is small but a domain expert is involved in the learning process by providing domain knowledge about constraints on the relevant network parameters. By incorporating qualitative parameter constraints with the limited data available, we provide an improved maximum a posteriori method.

The reminder of the paper is organized as follows: In Section 2, the related works on parameter learning are introduced, especially works with both sample data and parameter constraints. The BN parameter learning problem studied in this paper is formalized and described in detail in Section 3. The existing maximum a posteriori estimation algorithm, Qualitatively Maximum a Posteriori (QMAP) is introduced in Section 4, and further in Section 5, an improved qualitatively maximum a posteriori algorithm, Further Constrained Qualitatively Maximum a Posteriori (FC-QMAP) is presented. In Section 6, a theoretical analysis on constraint satisfaction for QMAP and FC-QMAP is presented. In Section 7, a detailed numerical example is given to illustrate the principle of the proposed FC-QMAP algorithm comparatively along with the ML and QMAP algorithms. In Section 8, a set of simulation experiments is presented with four typical benchmark Bayesian networks to compare the performance of the the proposed algorithm with other parameter learning algorithms. Finally, in Section 9, the main conclusion and findings of the paper are summarised, and several interesting future research directions are highlighted.

2. Related Work

For BN parameter learning from a small data set, the related research can be mainly grouped into two types: Non-MAP (Not Maximum a Posteriori) based methods and MAP (Maximum a Posteriori) based methods. For Non-MAP based methods, the BN parameters are computed by the optimization or the regulation of constrained parameter estimation models. Among those methods, Frank Wittig[14] proposed a constrained parameter learning algorithm. This algorithm defines the relative relations between a pair of parameters over two different distributions, and can be applied to cross-distribution parameter constraints[1]. First, a set of parameter constraint models is constructed from qualitative expert knowledge. Then, an optimization model consisting of an entropy function and the parameter constraints is built. Finally, the built optimization model is optimized using the Adaptive Probabilistic Networks method. Eric Altendorf[15] also discussed a parameter learning method applicable to cross-distribution constraints. What is interesting about this method

[1] Cross-distribution constraints are very common in real-world problems and are described as "monotonicity constraints", " order constraints" or " monotonic influence constraints " in the literature[14][15][20].

is that it defines an objective function that integrates the parameter constraint model into an entropy function, and the function is then solved by using gradient-descent algorithm. Ad Feelders[16] proposed the Isotonic Regression Estimation method concerned with cross-distribution constraints. Their algorithm employs the ML method to learn a set of initial parameters at the beginning, and then elicits parameter orders from parameter constraints. The initial parameters are regulated by the the algorithm so that the regulated parameters satisfy all the parameter orders. Takashi Isozaki[17] suggested the Minimum Free Energy method that is suitable for axiomatic parameter constraints[2]. Essentially, this method starts with constructing a free energy function. This energy function consists of the Kullback-Leibler divergence and an entropy function, and is used as the objective function. Furthermore the energy function and parameter constraints are integrated by Lagrange multipliers method, and gradient-decent method is employed to solve the problem. The Constrained Maximum Likelihood (CML) method was proposed by Cassio Campos[18]. This method works to any convex parameter constraints. Parameter constraints are transformed from domain expert knowledge, and a convex optimization model is then constructed with likelihood function and parameter constraints. The model can be optimized using a convex optimization method. Cassio Campos discussed the Constrained Maximum Entropy (CME) method applicable to any convex parameter constraints[19]. In this method, an Imprecise Dirichlet model that combines the prior information and the data set is created as a supplementary parameter constraint. Further, a convex optimization model containing an entropy function and convex parameter constraints is constructed. A convex optimization method can be applied to solve the formulated model. Yun Zhou[20] suggested a method for dealing with cross-distribution constraints, named Constrained Optimization with Flat Prior. An objective function is considered in this method that combines the likelihood function and the penalty function derived from constraint violations. The objective function is solved using sequential quadratic programming and the solutions are taken as the optimal parameters.

For the group of MAP-based methods, BN parameters are computed as linear interpolation values of the sample observations and the prior information. Among them, the Qualitative Maximum a Posteriori method[21] has been designed to tackle any convex parameter constraints. The method requires a certain amount of possible parameters to be sampled from parameter constraints using the rejection-acceptance sampling strategy. In addition, hyper-parameters of the prior Dirichlet distribution are determined as the products of a virtual sampling number and the mean values of the sampled parameters. The optimal parameters are then computed as the interpolations of the sample observations and hyper-parameters. A method, named as Bata Distribution Approximation-based Bayesian Estimation (BDABE) was presented in [22] to address intra-distribution parameter constraints[3]. Assuming the parameters of the prior distribution is a uniform distribution under parameter constraints, BDABE approximates the distribution using the beta distribution. The optimal parameters are further computed as the interpolation values of the sample observations and the prior parameters. Other methods for dealing with intra-distribution constraints include the Multi-nominal Parameter Learning with Constraints method[23]. This method counts the frequency of the configuration states of certain child and parent nodes, and then, an auxiliary BN model is built by integrating both the sample data and the parameter constraints. Furthermore, the optimal parameters are computed as the mean values of the probability distribution.

To the best of our knowledge, as a model-averaging method, the QMAP method is one of the best performed algorithms for parameter learning when the given data set is small. Essentially, QMAP estimation can be written as $\frac{N_{ijk}+M_{ijk}}{N_{ij}+M_{ij}}$, where N_{ij} and M_{ij} are the observation counts from the data set and the virtual data set, respectively, ~~where~~ node i is in the state k. N_{ijk} and M_{ijk} are the number of observations from the data set and the virtual data set in which node i is in the state k given the state j of its parent nodes. $\frac{N_{ijk}}{N_{ij}}$ and $\frac{M_{ijk}}{M_{ij}}$ are the estimation from data and the estimation from the parameter constraints, respectively. In this paper, we assume that the parameter constraints transformed from domain expert knowledge are all appropriate and correct, which means the estimation $\frac{M_{ijk}}{M_{ij}}$ satisfies all the parameter constraints regardless of the quality of the data. For the QMAP estimation, under the situation where the given data are extremely sparse or not informative, the data estimation $\frac{N_{ijk}}{N_{ij}}$ often violates the parameter constraints. In that case, the estimation $\frac{M_{ijk}}{M_{ij}}$ will be negatively influenced by $\frac{N_{ijk}}{N_{ij}}$, which makes the overall QMAP estimation violate part of the parameter constraints and thus fail to approach the true parameters. In an attempt to ensure that the overall QMAP estimating satisfies all the parameter constraints, we have developed this new algorithm FC-QMAP. The algorithm has the following features: (1) When the QMAP estimation satisfies all the parameter constraints, i.e., the QMAP estimation approaches the true parameter values well, the FC-QMAP estimation is equal to the QMAP estimation. (2) When the QMAP estimation violates any parameter constraints, the FC-QMAP estimation satisfies all the parameter constraints and thus approaches the true parameter values better than the QMAP estimation.

3. Preliminaries

3.1. Bayesian Network

A Bayesian network expresses the factorization of a joint probability distribution using an acyclic directed graph. A BN consists of its structure and parameters. Fig. 1 shows a typical Bayesian network, the brain tumor Bayesian network, where nodes in the network have the following meanings:

- $C \rightarrow Coma$
- $BT \rightarrow Brain\ Tumor$
- $SH \rightarrow Severe\ Headaches$
- $MC \rightarrow Metastatic\ Cancer$
- $CT \rightarrow Computed\ Tomography\ Scan$
- $ISC \rightarrow Increased\ Level\ of\ Serum\ Calcium$

[2]The axiomatic parameter constraints are from the law of probability thus are not required to be provided by the domain experts.

[3]The intra-distribution parameter constraints define the relations between parameters under one distribution. The constrained parameters also obey to the axiomatic parameter constraints

In the brain tumor Bayesian network, nodes like BT, MC and ISC, represent disease symptoms or diagnoses. Arrows from one node to another represent the influence of the node where an arrow starts to the node where the same arrow points to. Parameters like $P(C|BT, ISC)$ represent the strength of the joint influence imposed by the symptom nodes BT and ISC onto the diagnosis node C. In this paper, our objective is to learn the parameters of a Bayesian network, especially a discrete Bayesian network, whose structure is known beforehand, such as the brain tumor Bayesian network.

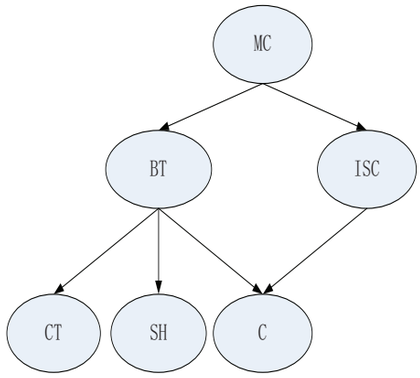

Figure 1: The brain tumor Bayesian network

3.2. Parameter Learning in a Bayesian Network

Learning parameters of a Bayesian Network entails estimating parameters from a given sample data set[4]. In this paper, samples with missing values are not considered. For a network with n node variables, parameter estimation can be expressed as a maximization problem of the log-likelihood function $\ell(\theta|D)$, where

$$\ell(\theta|D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} log \theta_{ijk} \tag{1}$$

According to the decomposability of BN, the parameter estimation of a network can be decomposed into the product of a set of independent estimations of individual variable nodes, and the Maximum Likelihood estimation of parameter θ_{ijk} is

$$\theta_{ijk} = \frac{N_{ijk}}{N_{ij}} \tag{2}$$

where r_i is the total state number of node i and $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

3.3. The Sample Complexity of Parameter Learning in a Fixed-structure Bayesian Network

When the given data is sufficient, the ML method is an ideal approach for accurate parameter learning. However, when the data set is small, ML estimation is often inaccurate. Therefore, defining the sample complexity[5] for BN parameter learning helps to determine whether ML is good enough to learn parameters with the expected accuracy. If not, more information such as parameter constraints is required. In relation to this, S. Dasgupta[24] has defined a way to calculate the sample complexity bounds for parameter learning with a fixed BN structure. For a network with n boolean nodes, if no node has more than k parents, with confidence $(1 - \delta)$, the sample complexity is upper-bounded by ~~the following:~~

$$\frac{288 n^2 2^k}{\epsilon^2} \ln^2(1 + \frac{3n}{\epsilon}) \ln \frac{1 + 3n/\epsilon}{\epsilon \delta} \tag{3}$$

where constant ϵ is the error rate and is often set as $\epsilon = \alpha n$~~,~~ for a small constant α.

3.4. Common Parameter Constraints

Generally, qualitative statements from domain experts can be translated into one of the following eight types of parameter constraints[15][25]:

(1) Axiomatic Constraint:

$$\sum_{k=1}^{r_i} \theta_{ijk} = 1, 0 \leq \theta_{ijk} \leq 1, \forall i, j, k \tag{4}$$

It describes the relation between parameters referring to the state j of a common parent configuration node with different states of a child node. This is a very basic constraint originating from the laws of probability.

(2) Range Constraint:

$$0 \leq \alpha_{ijk} \leq \theta_{ijk} \leq \beta_{ijk} \leq 1 \tag{5}$$

It defines the upper and lower bounds of a parameter and it is commonly used in reality. In addition, domain experts are comfortable with providing such constraints.

[4]In this paper, we focus on the learning of parameters in discrete Bayesian networks, which is also called "conditional probability table elicitation" in the literature.

[5]"Sample complexity" means the least number of samples required to learn parameters with a certain accuracy.

(3) Intra-distribution Constraint:

$$\theta_{ijk} \leq \theta_{ijk'}, \forall k \neq k' \tag{6}$$

It describes the comparative relation between two parameters referring to the same parent configuration node state j with different states k and k' of a child node.

(4) Cross-distribution Constraint:

$$\theta_{ijk} \leq \theta_{ij'k}, \forall j \neq j' \tag{7}$$

It defines the comparative relation between two parameters referring to the same child node state k with different parent configuration node states j and j'.

(5) Inter-distribution Constraint:

$$\theta_{ijk} \leq \theta_{i'j'k'}, \forall i \neq i', j \neq j', k \neq k' \tag{8}$$

It describes the comparative relation between two parameters referring to different nodes.

(6) Approximate-equality Constraint:

$$\theta_{ijk} \approx \theta_{i'j'k'}, \forall i \neq i', j \neq j', k \neq k' \tag{9}$$

It defines the close relation between any two parameters. Because the form of this constraint type is not convenient for further calculation, it needs to be transformed into the following form:

$$\mid \theta_{ijk} - \theta_{i'j'k'} \mid \leq \varepsilon, \forall i \neq i', j \neq j', k \neq k' \tag{10}$$

where ε is a very small value.

(7) Additive Synergy Constraint:

$$\theta_{ij_1k} + \theta_{ij_2k} \leq \theta_{ij_3k} + \theta_{ij_4k}, \forall i, k \tag{11}$$

It describes the comparative relation between the sums of two parameters referring to different configuration states of parent nodes.

(8) Product Synergy Constraint:

$$\theta_{ij_1k}\theta_{ij_2k} \leq \theta_{ij_3k}\theta_{ij_4k}, \forall i, k \tag{12}$$

It describes the comparative relation between products of two parameters referring to different parent configuration node states[6]. Note that the parameter constraints types 1-7 are all convex; whereas the product synergy constraint is non-convex.

4. Qualitatively Maximum a Posteriori Method

The *Qualitatively Maximum a Posteriori* method[21] is a posteriori estimation incorporating both quantitative data and qualitative constraints. The log-form score function of the method can be written as Eq. (13) and can be decomposed into data likelihood and parameter prior distribution.

$$logP(\theta|G, D, \Omega) = logP(D|\theta, G) + logP(\theta|\Omega, G) - c \tag{13}$$

where θ denotes the parameters of a network, G represents the fixed structure of the network, Ω is the set of parameter constraints and c is a constant and $c = P(D|\Omega, G)$.

The data likelihood equals to the conventional log-likelihood function as

$$logP(D|\theta, G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} log\theta_{ijk} \tag{14}$$

The prior distribution is defined by the parameter constraints, from which independent prior parameter instances can be sampled using the rejection-acceptance sampling method. Thus, the log-likelihood prior distribution can be written as

$$logP(\theta|\Omega, G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} M_{ijk} log\theta_{ijk} \tag{15}$$

where M_{ijk} is the pseudo prior statistic count for the ith node when it is in the the state k and its parent nodes are in the configuration state j.

As such, the overall QMAP log-likelihood score function can be written as

$$logP(\theta|G, D, \Omega) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} (N_{ijk} + M_{ijk}) log\theta_{ijk} - c \tag{16}$$

Finally, the maximum estimation of the QMAP score function are computed as

$$\widehat{\theta}_{ijk} = \frac{N_{ijk} + M_{ijk}}{\sum_{k=1}^{r_i}(N_{ijk} + M_{ijk})} = \frac{N_{ijk} + M_{ijk}}{N_{ij} + M_{ij}} \tag{17}$$

[6]The proposed method in this paper does not apply to Product Synergy Constraints, whose related knowledge is also very difficult for the domain experts to provide in reality.

where $M_{ijk} = A \cdot P(X_i = k, \Pi_i = j|\Omega)$. The best value of the coefficient A can be determined by cross-validation method. $P(X_i = k, \Pi_i = j|\Omega)$ is the mean value of the parameters sampled from the parameter constraints using the rejection-acceptance sampling method as

$$P(X_i = k, \Pi_i = j|\Omega) = \frac{\sum_{s=1}^{S} P_s(X_i = k, \Pi_i = j|\Omega)}{S} \tag{18}$$

where S is the number of sampled parameters, $P_s(X_i = k, \Pi_i = j|\Omega)$, satisfying all the known parameter constraints.

The *Qualitatively Maximum a Posteriori* method can be summarized as follows:

Step 1: Count the numbers of observations N_{ijk} and N_{ij} in the data set.

Step 2: Sample parameters that satisfy all the given constraints, $P_s(X_i = k, \Pi_i = j|\Omega), (s = 1, 2, 3...S)$[7] from the parameter constraints using the rejection-acceptance sampling method.

Step 3: Compute parameter value $P(X_i = k, \Pi_i = j|\Omega)$ according to Eq. (18).

Step 4: Determine the best coefficient A using cross-validation[8].

Step 5: Compute the QMAP estimation $\widehat{\theta}_{ijk}$ according to Eq. (17).

5. Further Constrained Qualitatively Maximum a Posteriori Method

The *Further Constrained Qualitatively Maximum a Posteriori* method is also a posteriori estimation method incorporating both quantitative data and qualitative constraints. The score function of FC-QMAP can be written as

$$P(\theta|G, D, \Omega) = P(D|\theta, G, \Omega)P(\theta|G, \Omega)/P(D|G, \Omega) \tag{19}$$

The log-form score function of FC-QMAP can be further written into

$$logP(\theta|G, D, \Omega) = logP(D|\theta, G, \Omega) + logP(\theta|G, \Omega) - logP(D|G, \Omega) \tag{20}$$

where, $logP(D|\theta, G, \Omega)$ is not the conventional log-likelihood function but a constrained log-likelihood function, which is

$$logP(D|\theta, G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} log\theta_{ijk} \tag{21}$$

$$Subject\ to\ \Omega(\theta_{ijk}) \leq 0 \tag{22}$$

$logP(\theta|G, \Omega)$ is defined in the same form as the QMAP method. Thus, the overall FC-QMAP log-likelihood score function can be written as

$$logP(\theta|G, D, \Omega) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} (N'_{ijk} + M_{ijk}) log\theta_{ijk} - logP(D|G, \Omega) \tag{23}$$

Finally, the maximum estimation of the FC-QMAP score function equals to

$$\widehat{\theta}_{ijk} = \frac{N'_{ijk} + M_{ijk}}{\sum_{k=1}^{r_i}(N'_{ijk} + M_{ijk})} = \frac{N'_{ijk} + M_{ijk}}{N_{ij} + M_{ij}} \tag{24}$$

where $N'_{ijk} = (\sum_{k=1}^{r_i} N_{ijk})\theta'_{ijk}$ and θ'_{ijk} are the optimization solutions of the model defined as:

$$\sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} log\theta_{ijk} \tag{25}$$

$$Subject\ to\ \Omega(\theta_{ijk}) \leq 0 \tag{26}$$

The *Further Constrained Qualitatively Maximum a Posteriori* method can be summarized as follows:

Step 1: Count the numbers of observations N_{ijk} and N_{ij} in the data set.

Step 2: Sample parameters that satisfy all the given constraints, $P_s(X_i = k, \Pi_i = j|\Omega), (s = 1, 2, 3...S)$ from the parameter constraints using the rejection-acceptance sampling method.

Step 3: Compute parameter value $P(X_i = k, \Pi_i = j|\Omega)$ according to Eq. (18).

Step 4: Determine the best coefficient A using cross-validation.

Step 5: Compute the QMAP estimation $\widehat{\theta}_{ijk}$ according to Eq. (17).

Step 6: If the QMAP estimation does not violate any parameter constraint, then stop the algorithm and the FC-QMAP estimation is equal to the QMAP estimation. If not, go to Step 7.

Step 7: Compute the parameter value θ'_{ijk} by optimizing the constrained likelihood model defined by Eqs. (25)-(26) using the convex optimization method[9].

Step 8: Compute the FC-QMAP estimation $\widehat{\theta}_{ijk}$ according to Eq. (24).

[7] The value of S in this paper is set to be 1000.

[8] In this paper, the best value of A is chosen from 1 to 20. Values larger than 20 are not considered because, when the value of A is too large, the prior statistic M_{ijk} and M_{ij} would dominate the whole QMAP or FC-QMAP estimation and make those estimations biased. Therefore, the accuracy of QMAP or FC-QMAP estimation would not improve even when amounts of data are imported.

[9] The Matlab software for convex optimization can be downloaded at http://cvxr.com/cvx/.

6. Theoretical Analysis of Constraint Satisfaction and Dissatisfaction

Axiom 1: All the convex parameter constraints about a certain parameter θ_{ijk} can be eventually transformed into an interval constraint $\theta_{ijk} \in [\theta^L_{ijk}, \theta^U_{ijk}]$.

Lemma 1: The Qualitatively Maximum a Posteriori estimation does not certainly guarantee the satisfaction of the parameter constraints.

Proof For Qualitatively Maximum a Posteriori estimation (Eq. (17)), to satisfy a known constraint, the following should hold

$$\theta^L_{ijk} \leq \frac{N_{ijk} + M_{ijk}}{N_{ij} + M_{ij}} \leq \theta^U_{ijk} \tag{27}$$

and this requires

$$(\theta^L_{ijk} N_{ij} + \theta^L_{ijk} M_{ij} - M_{ijk}) \leq N_{ijk} \leq (\theta^U_{ijk} N_{ij} + \theta^U_{ijk} M_{ij} - M_{ijk}) \tag{28}$$

However, for a data set of any size, the number of observations in the data set N_{ijk} could take any value that is larger than zero, i.e., $N_{ijk} \geq 0$. Therefore, when $N_{ijk} < (\theta^L_{ijk} N_{ij} + \theta^L_{ijk} M_{ij} - M_{ijk})$ holds or $N_{ijk} > (\theta^U_{ijk} N_{ij} + \theta^U_{ijk} M_{ij} - M_{ijk})$ holds, the QMAP estimation violates the parameter constraint.

■

Lemma 2: The Further Constrained Qualitatively Maximum a Posteriori estimation guarantees the satisfaction of all the parameter constraints.

Proof As the estimation θ'_{ijk} derived from the constrained likelihood model (Eqs. (25)-(26)) certainly satisfies all the parameter constraints, we have

$$\theta^L_{ijk} \leq \frac{N'_{ijk}}{N_{ij}} = \frac{N_{ij}\theta'_{ijk}}{N_{ij}} \leq \theta^U_{ijk} \tag{29}$$

and this means $\theta^L_{ijk} N_{ij} \leq N'_{ijk} \leq \theta^U_{ijk} N_{ij}$.

Meanwhile, since the mean value of the parameters sampled from the constraints, $P(X_i = k, \Pi_i = j|\Omega)$, also satisfies all the parameter constraints, we have

$$\theta^L_{ijk} \leq \frac{M_{ijk}}{M_{ij}} = \frac{A \cdot P(X_i = k, \Pi_i = j|\Omega)}{A \sum_{k=1}^{r_i} (P(X_i = k, \Pi_i = j|\Omega))} \leq \theta^U_{ijk} \tag{30}$$

which implies $\theta^L_{ijk} M_{ij} \leq M_{ijk} \leq \theta^U_{ijk} M_{ij}$.

Finally, we can derive $\theta^L_{ijk}(N_{ij} + M_{ij}) \leq (N'_{ijk} + M_{ijk}) \leq \theta^U_{ijk}(N_{ij} + M_{ij})$, and this is equivalent to

$$\theta^L_{ijk} \leq \frac{N'_{ijk} + M_{ijk}}{N_{ij} + M_{ij}} \leq \theta^U_{ijk} \tag{31}$$

Hence, the FC-QMAP estimation definitely satisfies all the parameter constraints.

■

7. Numerical Example

To illustrate the principle of the ML, QMAP and FC-QMAP methods, we consider an example Bayesian network extracted from the brain tumor Bayesian network shown in Section 3. We extract the following fragment of medical knowledge from [26]:

Consider a primary tumor with an uncertain prognosis with an patient. The cancer can metastasize to the brain and to any other organs. Metastatic cancer (MC) may be detected by an increased level of serum calcium (ISC). The presence of a brain tumor (BT) may be established from a CT scan (CT). Severe headaches (SH) are indicative of the presence of a brain tumor. Both a brain tumor and an increased level of serum calcium are likely to cause a patient to fall into a coma (C) in due course.

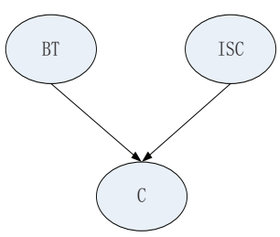

Figure 2: The example Bayesian network

For the sake of simplicity, we assign an index to each parameter in the example network as shown Table 1, and the parameter constraints transformed from the above medical expert knowledge are listed in Table 2.

Table 1: Parameter indexes

Index	Parameter	Index	Parameter
1	$P(C = 0 \mid BT = 0, ISC = 0)$	5	$P(C = 1 \mid BT = 0, ISC = 0)$
2	$P(C = 0 \mid BT = 0, ISC = 1)$	6	$P(C = 1 \mid BT = 0, ISC = 1)$
3	$P(C = 0 \mid BT = 1, ISC = 0)$	7	$P(C = 1 \mid BT = 1, ISC = 0)$
4	$P(C = 0 \mid BT = 1, ISC = 1)$	8	$P(C = 1 \mid BT = 1, ISC = 1)$

Table 2: Parameter constraints

Index	Constraint	Index	Constraint	Index	Constraint
1	$P_1 \geq P_5$	4	$P_8 \geq P_4$	7	$P_8 \geq P_7$
2	$P_6 \geq P_2$	5	$P_8 \geq P_5$	8	$P_6 \geq P_5$
3	$P_7 \geq P_3$	6	$P_8 \geq P_6$	9	$P_7 \geq P_5$

Then, we suppose that a small data set of 50 patient cases is available, from which the ML estimation, the QMAP estimation, and the FC-QMAP estimation can be computed as follows.

7.1. Maximum Likelihood Estimation

The ML estimation of the parameter $P(C \mid BT, ISC)$ is computed as follows:

$$\begin{array}{ll} P(C=0 \mid BT=0, ISC=0)=\frac{5}{13}=0.385, & P(C=0 \mid BT=0, ISC=1)=\frac{3}{12}=0.250 \\ P(C=0 \mid BT=1, ISC=0)=\frac{6}{13}=0.462, & P(C=0 \mid BT=1, ISC=1)=\frac{4}{12}=0.333 \\ P(C=1 \mid BT=0, ISC=0)=\frac{8}{13}=0.615, & P(C=1 \mid BT=0, ISC=1)=\frac{9}{12}=0.750 \\ P(C=1 \mid BT=1, ISC=0)=\frac{7}{13}=0.538, & P(C=1 \mid BT=1, ISC=1)=\frac{8}{12}=0.667 \end{array}$$

Note that:
(1) $P(C=1 \mid BT=0, ISC=0) \geq P(C=0 \mid BT=0, ISC=0)$ implies that a person without an increased serum calcium level and with a brain tumor would have only a 38.5% chance of staying healthy (not falling into a coma) and have a 61.5% chance of falling into a coma.
(2) $P(C=1 \mid BT=0, ISC=1) \geq P(C=1 \mid BT=1, ISC=1)$ implies that a person with an increased serum calcium level but no brain tumor would have a 75% chance of falling into a coma; whereas a person with an increased serum calcium level as well as a brain tumor would see this probability dropped to 66.7%.
(3) $P(C=1 \mid BT=0, ISC=0) \geq P(C=1 \mid BT=1, ISC=0)$ implies that a person without an increased serum calcium level and no brain tumor would have a 61.5% chance of falling into a coma; whereas a person without an increased serum calcium level but with a brain tumor would have this probability dropped to 53.8%.
It is evident that the ML estimation fails to satisfy all the parameter constraints. Thus, a network from which the above counterintuitive inferences have been made would not be well accepted by any physician.

7.2. Qualitatively Maximum a Posteriori Estimation

For the given data set, the best value of the coefficient A can be found to be 4 via cross-validation. The QMAP estimation of $P(C \mid B, ISC)$ is computed as follows:

$$\begin{array}{ll} P(C=0 \mid BT=0, ISC=0)=\frac{4\times 0.751+5}{4+13}=0.471, & P(C=0 \mid BT=0, ISC=1)=\frac{4\times 0.315+3}{4+12}=0.266 \\ P(C=0 \mid BT=1, ISC=0)=\frac{4\times 0.315+6}{4+13}=0.427, & P(C=0 \mid BT=1, ISC=1)=\frac{4\times 0.124+4}{4+12}=0.281 \\ P(C=1 \mid BT=0, ISC=0)=\frac{4\times 0.249+8}{4+13}=0.529, & P(C=1 \mid BT=0, ISC=1)=\frac{4\times 0.685+9}{4+12}=0.734 \\ P(C=1 \mid BT=1, ISC=0)=\frac{4\times 0.685+7}{4+13}=0.573, & P(C=1 \mid BT=1, ISC=1)=\frac{4\times 0.876+8}{4+12}=0.719 \end{array}$$

and the results can be summarised as
(1) $P(C=1 \mid BT=0, ISC=0) \geq P(C=0 \mid BT=0, ISC=0)$ implies that a person without an increased serum calcium level and with a brain tumor would have only a 47.1% chance of staying healthy (not falling into a coma) and 52.9% chance of falling into a coma.
(2) $P(C=1 \mid BT=0, ISC=1) \geq P(C=1 \mid BT=1, ISC=1)$ implies that a patient with an increased serum calcium level but no brain tumor would have a 73.4% chance of falling into a coma; whereas a patient with an increased serum calcium level and a brain tumor would have this probability dropped to 71.9%.
Finally, we can find that the QMAP estimation also fails to satisfy all the parameter constraints.

7.3. Further Constrained Qualitatively Maximum a Posteriori Estimation

Based on the FC-QMAP method, the FC-QMAP estimation of the parameter $P(C \mid BT, ISC)$ is computed as follows:

$$\begin{array}{ll} P(C=0 \mid BT=0, ISC=0)=\frac{4\times 0.751+13\times 0.500}{4+13}=0.559, & P(C=0 \mid BT=0, ISC=1)=\frac{4\times 0.315+12\times 0.292}{4+12}=0.298 \\ P(C=0 \mid BT=1, ISC=0)=\frac{4\times 0.315+13\times 0.462}{4+13}=0.427, & P(C=0 \mid BT=1, ISC=1)=\frac{4\times 0.124+12\times 0.292}{4+12}=0.250 \\ P(C=1 \mid BT=0, ISC=0)=\frac{4\times 0.249+13\times 0.500}{4+13}=0.441, & P(C=1 \mid BT=0, ISC=1)=\frac{4\times 0.685+12\times 0.708}{4+12}=0.702 \\ P(C=1 \mid BT=1, ISC=0)=\frac{4\times 0.685+13\times 0.538}{4+13}=0.573, & P(C=1 \mid BT=1, ISC=1)=\frac{4\times 0.876+12\times 0.708}{4+12}=0.750 \end{array}$$

and it can be found that the FC-QMAP estimation satisfies all the parameter constraints.

8. Experiments

We further carry out experiments to evaluate the learning performance of different algorithms in terms of learning accuracy and learning efficiency. The learning accuracy is evaluated by the Kullback-Leibler (KL) divergence[27], which indicates the divergence between the learnt parameters and the true parameters. The learning efficiency is evaluated by the learning time. In the following experiments, we consider seven algorithms: Maximum Likelihood, Constrained Maximum Likelihood, Maximum Entropy, Constrained Maximum Entropy, Maximum a Posteriori, Qualitatively Maximum a Posteriori and Further Constrained Qualitatively Maximum a Posteriori [10]. We perform experiments on four publicly

[10]The codes are available upon request

available benchmark networks, and the size of these networks varies from small, medium, large to very large. The summary of the networks is given in Table 3.

Note that: (1) For the MAP method, flat prior is considered in the experiments. (2) For the QMAP and the FC-QMAP methods, the best value of the coefficient A is initially chosen as between 1 and 20 and is then determined by cross-validation.

Table 3: Details of Bayesian networks in the experiments

	Asia	Alarm	Win95pts	Andes
Nodes	8	37	76	223
Edges	8	46	112	338
Parameters	18	509	574	1157
Size classification	Small	Medium	Large	Very Large

8.1. Parameter learning with different sample sizes

First of all, we look at the learning performance of different algorithms with different sample sizes.

Experiment settings: (1) The data sparsity considered in this experiment is: 50, 100, 150 and 200, respectively. (2) The parameter constraints are generated from the true parameters of the networks and the maximum number of constraints for each node is 30. The type of generated constraints varies from the axiomatic constraints, the range constraints, the approximate-equality constraints, the intra-distribution constraints to the cross-distribution constraints. These types of constraints are all common in real-world problems. The average KL divergence and running time for different networks are summarized in Table 4 and Table 5, where the best results are highlighted in bold.

Table 4: Average KL divergence of different algorithms under different sample sizes

	ML	CML	ME	CME	MAP	QMAP	FC-QMAP
(a) Asia network							
50	0.543±0.321	0.063±0.024	0.197±0.015	0.134±0.027	0.021±0.005	0.009±0.001	**0.007±0.001**
100	0.372±0.191	0.064±0.022	0.168±0.012	0.107±0.046	0.017±0.006	0.008±0.002	**0.006±0.001**
150	0.333±0.181	0.069±0.022	0.147±0.012	0.084±0.043	0.015±0.006	0.007±0.001	**0.006±0.001**
200	0.207±0.144	0.061±0.018	0.129±0.013	0.061±0.015	0.013±0.006	0.006±0.001	**0.005±0.001**
(b) Alarm network							
50	0.502±0.061	0.225±0.022	0.264±0.007	0.129±0.004	0.214±0.016	0.124±0.002	**0.122±0.002**
100	0.485±0.046	0.231±0.027	0.232±0.008	0.113±0.004	0.181±0.008	0.114±0.002	**0.113±0.002**
150	0.436±0.043	0.212±0.024	0.212±0.006	**0.105±0.004**	0.168±0.009	0.109±0.003	0.108±0.003
200	0.429±0.054	0.211±0.025	0.195±0.005	**0.098±0.003**	0.157±0.008	0.103±0.002	0.102±0.002
(c) Win95pts network							
50	0.717±0.092	0.229±0.020	0.262±0.002	0.142±0.001	0.237±0.013	0.129±0.001	**0.128±0.001**
100	0.658±0.095	0.211±0.029	0.244±0.003	0.137±0.001	0.221±0.014	**0.125±0.001**	**0.125±0.001**
150	0.605±0.095	0.197±0.023	0.231±0.003	0.134±0.001	0.215±0.013	0.123±0.001	**0.122±0.001**
200	0.604±0.085	0.179±0.027	0.222±0.001	0.132±0.001	0.216±0.011	0.121±0.001	**0.119±0.001**
(d) Andes network							
50	1.072±0.082	0.254±0.042	0.221±0.002	0.078±0.001	0.195±0.011	0.049±0.001	**0.046±0.001**
100	0.861±0.087	0.249±0.034	0.173±0.003	0.067±0.002	0.153±0.013	0.044±0.001	**0.041±0.001**
150	0.763±0.062	0.242±0.044	0.147±0.002	0.059±0.001	0.133±0.011	0.041±0.001	**0.037±0.001**
200	0.694±0.064	0.233±0.031	0.129±0.003	0.053±0.001	0.120±0.011	0.037±0.001	**0.034±0.001**

Learning accuracy analysis: (1) With the increase of data size, the performance of all the algorithms improves by different degrees. (2) In almost all the cases, FC-QMAP outperforms the other learning algorithms. (3) With the increase of data, the QMAP estimation gradually approaches the FC-QMAP estimation. A reasonable explanation for this would be that, with more data, the overall QMAP estimation becomes less likely to violate the constraints, and this means less regulations are required.

Table 5: Running time (seconds) of different algorithms under different sample sizes

	ML	CML	ME	CME	MAP	QMAP	FC-QMAP
(a) Asia network							
50	**0.001±0.001**	1.056±0.026	0.881±0.045	1.115±0.079	**0.001±0.001**	0.148±0.016	0.507±0.175
100	**0.002±0.001**	1.081±0.025	0.896±0.073	1.130±0.049	**0.002±0.001**	0.255±0.017	0.578±0.199
150	**0.002±0.001**	1.099±0.027	0.927±0.072	1.143±0.132	**0.002±0.001**	0.383±0.049	0.665±0.206
200	**0.003±0.001**	1.145±0.039	0.885±0.059	1.149±0.071	**0.003±0.001**	0.489±0.038	0.682±0.194
(b) Alarm network							
50	**0.001±0.001**	0.315±0.007	0.335±0.098	0.807±0.366	**0.001±0.001**	0.122±0.043	0.232±0.075
100	**0.001±0.001**	0.333±0.008	0.363±0.143	0.489±0.177	**0.001±0.001**	0.246±0.108	0.369±0.154
150	**0.002±0.001**	0.346±0.009	0.359±0.146	0.584±0.314	**0.002±0.001**	0.356±0.165	0.495±0.222
200	**0.003±0.001**	0.356±0.005	0.369±0.134	0.503±0.184	**0.003±0.001**	0.479±0.197	0.631±0.276
(c) Win95pts network							
50	**0.001±0.001**	0.319±0.011	0.344±0.001	0.429±0.013	**0.001±0.001**	0.095±0.002	0.138±0.017
100	**0.002±0.001**	0.331±0.016	0.346±0.001	0.439±0.011	**0.002±0.001**	0.173±0.002	0.229±0.019
150	**0.002±0.001**	0.335±0.006	0.352±0.009	0.437±0.015	**0.002±0.001**	0.253±0.004	0.319±0.013
200	**0.002±0.001**	0.341±0.015	0.356±0.014	0.430±0.014	**0.002±0.001**	0.333±0.008	0.403±0.015
(d) Andes network							
50	**0.002±0.001**	0.482±0.058	0.415±0.061	0.511±0.087	**0.002±0.001**	0.096±0.007	0.351±0.047
100	**0.002±0.001**	0.515±0.061	0.414±0.049	0.523±0.081	**0.002±0.001**	0.176±0.012	0.455±0.059
150	**0.002±0.001**	0.557±0.132	0.435±0.104	0.559±0.151	**0.002±0.001**	0.259±0.029	0.549±0.127
200	**0.003±0.001**	0.541±0.082	0.418±0.064	0.533±0.098	**0.003±0.001**	0.333±0.026	0.595±0.077

Learning efficiency analysis: (1) With the increase of data, the time consumption of almost all the algorithms increases but does not change significantly. (2) The FC-QMAP method is more time-consuming than the QMAP method

since the FC-QMAP method occasionally employs the convex optimization approach for further optimization to ensure that the estimation satisfies all the constraints.

8.2. Parameter learning with different constraint sizes

We further explore the learning performance of different learning algorithms with different constraint sizes.

Experiment settings: (1) The size of all the data sets for all the test networks is set to be 50, which is a small number for each network. (2) The parameter constraints are generated from the true parameters of the networks and the maximum number of constraints for each node is 30. We learn parameters with fixed data set and increasing the amount of parameter constraints chosen from the generated constraints. The constraint sparsity varies from 25% to 100%. The average KL divergence and the running time of different networks are summarized in Table 6 and Table 7, and the best results are highlighted in bold.

Table 6: Average KL divergence of different algorithms under different constraint sizes

	ML	CML	ME	CME	MAP	QMAP	FC-QMAP
(a) Asia network							
25%	0.448±0.139	0.311±0.169	0.286±0.008	0.156±0.031	0.134±0.037	0.110±0.029	**0.107±0.028**
50%	0.448±0.139	0.079±0.041	0.286±0.008	0.066±0.029	0.134±0.037	0.046±0.025	**0.041±0.025**
75%	0.448±0.139	0.042±0.025	0.286±0.008	0.027±0.017	0.134±0.037	0.015±0.007	**0.009±0.007**
100%	0.448±0.139	0.022±0.010	0.286±0.008	0.009±0.002	0.134±0.037	0.009±0.002	**0.002±0.001**
(b) Alarm network							
25%	0.518±0.094	0.404±0.068	0.328±0.003	0.263±0.005	**0.215±0.012**	0.255±0.008	0.254±0.007
50%	0.518±0.094	0.322±0.061	0.328±0.003	0.208±0.007	0.215±0.012	0.197±0.007	**0.196±0.007**
75%	0.518±0.094	0.239±0.047	0.328±0.003	0.169±0.007	0.215±0.012	0.160±0.007	**0.158±0.007**
100%	0.518±0.094	0.193±0.047	0.328±0.003	0.138±0.002	0.215±0.012	0.128±0.002	**0.126±0.002**
(c) Win95pts network							
25%	0.538±0.116	0.382±0.082	0.299±0.001	0.233±0.005	0.233±0.008	0.203±0.005	**0.203±0.005***
50%	0.538±0.116	0.281±0.051	0.299±0.001	0.184±0.005	0.233±0.008	0.162±0.005	**0.162±0.005***
75%	0.538±0.116	0.198±0.021	0.299±0.001	0.153±0.004	0.233±0.008	0.138±0.003	**0.138±0.003***
100%	0.538±0.116	0.156±0.015	0.299±0.001	0.139±0.001	0.233±0.008	0.126±0.001	**0.126±0.001***
(d) Andes network							
25%	1.040±0.052	0.707±0.045	0.316±0.001	0.209±0.003	0.189±0.008	0.144±0.002	**0.142±0.002**
50%	1.040±0.052	0.480±0.052	0.316±0.001	0.139±0.003	0.189±0.008	0.096±0.002	**0.093±0.002**
75%	1.040±0.052	0.345±0.046	0.316±0.001	0.094±0.001	0.189±0.008	0.064±0.002	**0.060±0.002**
100%	1.040±0.052	0.252±0.042	0.316±0.001	0.071±0.001	0.189±0.008	0.047±0.001	**0.042±0.001**

* The FC-QMAP estimation is slightly more accurate than the QMAP estimation. And the detailed KL divergence are: QMAP (0.2027±0.0050, 0.1620±0.0048, 0.1384±0.0032, 0.1262±0.0007), FC-QMAP (0.2025±0.0050, 0.1616±0.0048, 0.1377±0.0031, 0.1255±0.0006).

Learning accuracy analysis: (1) For the algorithms not using constraints, such as ML, ME, MAP, changing the constraint size has no impact on their performance; However, for the algorithms incorporating constraints, such as CML, CME, QMAP, FC-QMAP, the increase of constraints affects their performance by a certain degree depending on the number of constraints they work with. (2) In most the cases, FC-QMAP outperforms other parameter learning algorithms except MAP on the Alarm network when the number of parameter constraints is limited (only 25% of the generated constraints are considered). (3) Compared with the QMAP, the FC-QMAP performs better especially when more constraints are available. This suggests that, with the increase of constraints, the QMAP estimation is more likely to violate the parameter constraints.

Table 7: Running time (seconds) of different algorithms under different constraint sizes

	ML	CML	ME	CME	MAP	QMAP	FC-QMAP
(a) Asia network							
25%	**0.001±0.001**	1.001±0.058	0.821±0.054	0.967±0.062	**0.001±0.001**	0.153±0.009	0.507±0.164
50%	**0.001±0.001**	1.086±0.071	0.814±0.032	1.057±0.066	**0.001±0.001**	0.148±0.016	0.699±0.186
75%	**0.001±0.001**	1.175±0.055	0.800±0.046	1.164±0.092	**0.001±0.001**	0.151±0.016	0.842±0.089
100%	**0.001±0.001**	1.331±0.079	0.800±0.065	1.272±0.068	**0.001±0.001**	0.146±0.007	0.929±0.123
(b) Alarm network							
25%	**0.001±0.001**	0.233±0.052	0.266±0.058	0.352±0.015	**0.001±0.001**	0.112±0.030	0.297±0.029
50%	**0.001±0.001**	0.284±0.083	0.281±0.058	0.425±0.136	**0.001±0.001**	0.118±0.044	0.315±0.109
75%	**0.001±0.001**	0.351±0.092	0.294±0.074	0.503±0.125	**0.001±0.001**	0.121±0.035	0.398±0.098
100%	**0.001±0.001**	0.419±0.089	0.294±0.063	0.566±0.119	**0.001±0.001**	0.122±0.031	0.472±0.098
(c) Win95pts network							
25%	**0.001±0.001**	0.275±0.003	0.329±0.009	0.381±0.007	**0.001±0.001**	0.094±0.001	0.156±0.009
50%	**0.001±0.001**	0.313±0.005	0.331±0.010	0.419±0.006	**0.001±0.001**	0.094±0.001	0.201±0.017
75%	**0.001±0.001**	0.346±0.005	0.324±0.004	0.461±0.012	**0.001±0.001**	0.094±0.001	0.241±0.023
100%	**0.001±0.001**	0.406±0.015	0.324±0.008	0.521±0.011	**0.001±0.001**	0.094±0.002	0.281±0.031
(d) Andes network							
25%	**0.002±0.001**	0.422±0.046	0.368±0.047	0.458±0.046	**0.002±0.001**	0.095±0.007	0.359±0.037
50%	**0.002±0.001**	0.465±0.051	0.368±0.045	0.502±0.051	**0.002±0.001**	0.095±0.006	0.462±0.043
75%	**0.002±0.001**	0.546±0.111	0.387±0.075	0.619±0.111	**0.002±0.001**	0.098±0.012	0.570±0.117
100%	**0.002±0.001**	0.601±0.099	0.376±0.057	0.673±0.099	**0.002±0.001**	0.096±0.008	0.626±0.103

Learning efficiency analysis: (1) With the increase of the constraints, the time consumption of the algorithms not considering constraints does not change notably, and the time consumption of most of the algorithms (except the QMAP) working with constraints increases by different degrees. (2) The ML and the MAP are still the most efficient methods since they are all analytical solutions. (3) The FC-QMAP is less efficient than the QMAP, especially when the number of constraints increases. This can be interpreted as, when the constraint number increases, the QMAP estimation is more likely to violate the constraints and the FC-QMAP would regulate the QMAP estimation by the convex optimization, which is time-consuming.

8.3. Parameter learning with different error constraint sizes

As shown in the first two experiments, incorporating parameter constraints can improve the learning performance, especially when the given data is sparse. However, in practice, parameter constraints transformed from the domain knowledge are sometimes inadequate and incorrect. To deal with that situation, we consider error constraints in the learning process to testify the robustness of the estimators.

Experiment settings: (1) The size of the data sets for all the testing networks is set to be 50. (2) The parameter constraints are generated from the networks and the maximum number of constraints for each node is 30. Since domain knowledge with significant error would not be taken into account, therefore only a small amount of error constraints, varying from 2 to 10, is incorporated into the parameter learning process. The averaged KL divergence for different networks under different error constraint sizes is summarized in Table 8, and the best results are highlighted in bold.

Table 8: Averaged KL divergence of different algorithms under different error constraint sizes

	ML	CML	ME	CME	MAP	QMAP	FC-QMAP
(a) Asia network							
N_{error}=2	0.689±0.418	0.290±0.119	0.283±0.003	0.209±0.010	**0.135±0.041**	0.138±0.037	0.149±0.028
N_{error}=4	0.689±0.418	0.388±0.215	0.283±0.003	0.217±0.009	**0.135±0.041**	0.201±0.031	0.227±0.029
N_{error}=6	0.689±0.418	0.545±0.359	0.283±0.003	0.229±0.005	**0.135±0.041**	0.258±0.023	0.297±0.027
P_{error}=8	0.689±0.418	0.548±0.389	0.283±0.003	0.239±0.005	**0.135±0.041**	0.289±0.036	0.371±0.038
N_{error}=10	0.689±0.418	0.645±0.399	0.283±0.003	0.246±0.003	**0.135±0.041**	0.347±0.034	0.454±0.037
(b) Alarm network							
N_{error}=2	0.543±0.056	0.319±0.023	0.329±0.004	0.265±0.034	0.224±0.018	**0.173±0.012**	0.174±0.030
N_{error}=4	0.543±0.056	0.322±0.034	0.329±0.004	0.277±0.028	0.224±0.018	**0.178±0.014**	0.181±0.043
N_{error}=6	0.543±0.056	0.343±0.037	0.329±0.004	0.289±0.052	0.224±0.018	**0.184±0.015**	0.187±0.024
N_{error}=8	0.543±0.056	0.361±0.053	0.329±0.004	0.304±0.045	0.224±0.018	**0.190±0.010**	0.195±0.034
N_{error}=10	0.543±0.056	0.379±0.099	0.329±0.004	0.320±0.089	0.224±0.018	**0.194±0.018**	0.198±0.043
(c) Win95pts network							
N_{error}=2	0.550±0.099	0.436±0.073	0.299±0.001	0.265±0.001	0.228±0.009	**0.175±0.001**	0.177±0.001
N_{error}=4	0.550±0.099	0.437±0.072	0.299±0.001	0.266±0.001	0.228±0.009	**0.176±0.001**	0.180±0.001
N_{error}=6	0.550±0.099	0.438±0.072	0.299±0.001	0.266±0.001	0.228±0.009	**0.177±0.001**	0.183±0.001
N_{error}=8	0.550±0.099	0.438±0.072	0.299±0.001	0.266±0.001	0.228±0.009	**0.178±0.001**	0.185±0.001
N_{error}=10	0.550±0.099	0.438±0.071	0.299±0.001	0.267±0.001	0.228±0.009	**0.179±0.001**	0.188±0.001
(d) Andes network							
N_{error}=2	1.082±0.139	0.671±0.023	0.316±0.001	0.217±0.001	0.197±0.017	**0.071±0.001***	0.071±0.001
N_{error}=4	1.082±0.139	0.674±0.033	0.316±0.001	0.217±0.001	0.197±0.017	**0.071±0.001**	0.072±0.001
N_{error}=6	1.082±0.139	0.672±0.029	0.316±0.001	0.218±0.001	0.197±0.017	**0.072±0.001**	0.073±0.001
N_{error}=8	1.082±0.139	0.671±0.028	0.316±0.001	0.218±0.001	0.197±0.017	**0.072±0.001**	0.074±0.001
N_{error}=10	1.082±0.139	0.670±0.028	0.316±0.001	0.218±0.001	0.197±0.017	**0.074±0.001**	0.075±0.001

* The QMAP estimation is slightly more accurate than the FC-QMAP estimation. And the detailed KL divergence are: QMAP (0.0710±0.0005), FC-QMAP (0.0712±0.0005).

Learning accuracy analysis: (1) For the algorithms not using constraints, such as ML, ME, MAP, error constraints have no impact on their performance. On the other hand, for the algorithms considering constraints, such as CML, CME, QMAP, FC-QMAP, error constraints negatively affect their performance, and the degree of the influence depends on the size of the networks they work on and the number of error constraints involved. (2) For the fixed amount of error constraints, the parameter learning of a larger network, such as Win95pts and Andes network, is more robust than that of small networks. (3) Compared with the QMAP, the FC-QMAP is less robust and more sensitive to the accuracy of the parameter constraints. Therefore, before applying the FC-QMAP for parameter learning, it is advisable to verify and validate the parameter constraints or domain knowledge.

9. Conclusion

The existing advanced parameter learning algorithms, such as the QMAP algorithm, fail to satisfy all the parameter constraints, especially when only limited data is available. In this paper, we have presented an adaptive QMAP method, the FC-QMAP method in an attempt to ensure that the QMAP estimation satisfies all the parameter constraints. The proposed method has the following features:

Advantages: It is guaranteed that the learnt parameters satisfy all the convex parameter constraints under any cases, and this is the main improvement to the original QMAP method.

Disadvantages: (1) The proposed method does not apply to non-convex parameter constraints. (2) The proposed method is more time-consuming than some of the existing algorithms.

In the future, the relevant work will focus on improving the algorithm's performance, including: (1) Improving the applicability of the proposed method to include non-convex parameter constraints. When non-convex parameter constraints are imported, we can further adjust FC-QMAP estimation by methods such as Isotonic Regression. (2) Improving the efficiency of the proposed method. When the QMAP estimation violates only range constraints or intra-distribution parameter constraints, we can replace the optimization of the constrained likelihood function by a constrained linear function, which can be efficiently optimized by the linear programming technology.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (Under grant number 61573285) and the Innovation Foundation for Doctor Dissertations of Northwestern Polytechnical University (Under grant number CX201619).

References

[1] Pearl, J.: Probabilistic reasoning in intelligent systems. Massachusetts: Morgan Kaufmann. (1988)

[2] Tamda, Y., Imoto, S.,Araki, H.: Estimating genome-wide gene networks using nonparametric Bayesian network models on massively parallel computers. IEEE Transactions on Computational Biology and Bioinformatics. 3, 683-697 (2011)

[3] Infantes, G., Ghallab, M.,Ingrand, F.: Learning the behavior model of a robot. Auton Robot. 30, 157-177 (2011)

[4] Ibrahim, W., Beiu, V.: Using bayesian networks to accurately calculate the reliability of complementary metal oxide semiconductor gates. IEEE Transactions on Reliability. 60, 538-549 (2011)

[5] Steven, M., Ann, N.,Kevin, K.: Anomaly detection in vessel tracks using Bayesian networks. International Journal of Approximate Reasoning. 55, 84-98 (2014)

[6] Neil, W., Mahmood, R.: Detection and classification of non-stationary transient signals using sparse approximations and Bayesian networks. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 22, 1750-1764 (2014)

[7] Dries, L., Steven, B., Rob, D.,Guy, E., Joris, A.: A review of Bayesian belief networks in ecosystem service modeling. Environmental Modeling & Software. 46, 1-11 (2013)

[8] Almond, R., Mislevy, R., Steinburg, L.,Yan, D., Williamson, D.: Bayesian Networks in Educational Assessment. New York: Springer. (2015)

[9] Redner, R., Walker, H.: Mixture densities, maximum likelihood and the EM algorithm. SIAM Review. 26, 195-239 (1984)

[10] Seixas, F., Zadrozny, B., Laks, J, Conci, A., Saade, D.: A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment. Computers in Biology and Medicine. 51, 140-158 (2014)

[11] Ji-lei, Hu, Xiao-wei Tang, Jiang-nan Qiu: A Bayesian network approach for predicting seismic liquefaction based on interpretive structural modeling. Georisk. 9(3), 200-217 (2015)

[12] Constantinou, A., Freestone, M., Marsh, W., Fenton, N., Coid, J.: Risk assessment and risk management of violent reoffending among prisoners. Expert Systems with Applications. 42(21), 7511-7529 (2015)

[13] Helsper, E., Gaag, L., Groenendal, F.: Designing a procedure for the acquisition of probability constraints for Bayesian networks. Proceedings of the Fourteenth Conference on Engineering Knowledge in the Age of the Semantic Web. 280-292 (2004)

[14] Wittig, F., Jameson, A.: Exploiting qualitative knowledge in the learning of conditional probabilities of Bayesian networks. Proceedings of the Sixteenth International Conference on Uncertainty in Artificial Intelligence. 644-652 (2000)

[15] Altendorf, E., Restificar, A., Dietterich, T.: Learning from sparse data by exploiting monotonicity constraints. Proceedings of the Twenty First International Conference on Uncertainty in Artificial Intelligence. 18-26 (2005)

[16] Feelders, A., Gaag, L.: Learning bayesian networks parameters under order constraints. International Journal of Approximate Reasoning. 42, 37-53 (2006)

[17] Isozaki, T., Kato, N., Ueno, M.: Data temperature in minimum free energies for parameter learning of Bayesian networks. International Journal on Artificial Intelligence Tools. 18, 653-671 (2009)

[18] Campos, C., Yan, T., Ji, Q.: Constrained maximum likelihood learning of Bayesian networks for facial action recognition. Proceeding of the Tenth European Conference on Computer Vision. 168-181 (2008)

[19] Campos, C., Ji, Q.: Improving Bayesian network parameter learning using constraints. Proceedings of the Nineteenth International Conference on Pattern Recognition. 1-4 (2008)

[20] Yun Z, Norman Fenton, Cheng Z.: An empirical study of Bayesian network parameter learning with monotonic influence constraints. Decision Support Systems. 87, 69-79 (2016)

[21] Rui, C., Wei, W.: Novel algorithm for Bayesian network parameter learning with informative prior constraints. International Joint Conference on Neural Networks. 1-8 (2010)

[22] Ruo-hai, D., Xiao-guang, G., Zhi-gao, G.: Discrete Bayesian network parameter learning based on monotonic constraint. Journal of Systems Engineering and Electronics. 36, 272-277 (2014)

[23] Yun, Z., Fenton, N., Neil, M.: Bayesian network approach to multinomial parameter learning using data and expert judgments. International Journal of Approximate Reasoning. 55, 1252-1268 (2014)

[24] Dasgupta, S.: The sample complexity of learning fixed-structure Bayesian Networks. Machine Learning. 29, 165-180 (1997)

[25] Wellman, M.: Fundamental concepts of qualitative probabilistic networks. Artificial Intelligence. 44(3): 257-303(1990)

[26] Cooper, G.: NESTOR: a computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. Stanford University. (1984).

[27] Kullback, S., Leibler, R.: On information and sufficiency. Annals of Mathematical Statistics. 22, 79-86 (1951)